풀이

STEP I 2005, 문제 13

확률과 통계 • 대칭 분포

> **STEP I 2005, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/05-s1-q13

핵심 아이디어

분포(distribution)의 모양에 대해서는 두 개의 누적확률(cumulative probability)과 대칭성(symmetry) 외에는 아무것도 주어지지 않으므로, 모든 것을 이것들만으로 구성해야 한다. μ에 대한 대칭성으로부터 두 가지 사실을 덤으로 얻는다. 확률의 절반은 μ의 양쪽에 각각 놓이고, $\mu - k\sigma$ 아래의 꼬리(tail)는 $\mu + k\sigma$ 위의 꼬리와 같다. 소문항 (ii)의 요령은 언급된 모든 부피(495, 500, 505 ml)가 두 종류의 우유 중 어느 한쪽에 대하여 μ, $\mu \pm \frac{1}{2}\sigma$ 또는 $\mu + \sigma$에 놓인다는 것을 알아차리는 것이다.

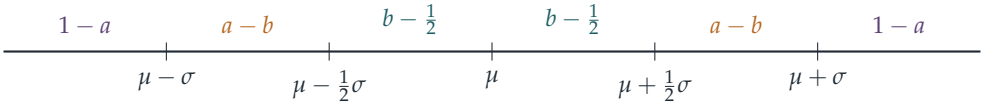
풀이

X는 부피를 모형화하므로 줄곧 연속(continuous) 확률변수로 다룬다. 따라서 한 값의 확률은 0이고, $<$로 쓰든 $\leq$로 쓰든 상관없다.

μ에 대한 대칭성으로부터 $\mathrm{P}(X \leq \mu) = \frac{1}{2}$이고, 임의의 k에 대하여 $\mathrm{P}(X \leq \mu - k\sigma) = \mathrm{P}(X \geq \mu + k\sigma)$이다. 특히

$$\mathrm{P}(X \leq \mu - \tfrac{1}{2}\sigma) = 1 - b, \qquad \mathrm{P}(X \leq \mu - \sigma) = 1 - a.$$

μ로부터 σ 단위로 잰 각 구간 띠의 확률을 적어 두면 도움이 된다.



예를 들어 $\mathrm{P}(\mu < X \leq \mu + \frac{1}{2}\sigma) = b - \frac{1}{2}$이고 $\mathrm{P}(\mu + \frac{1}{2}\sigma < X \leq \mu + \sigma) = a - b$이다. 왼쪽 띠들은 대칭성으로부터 얻어진다.

소문항 (i)

누적확률을 빼면

$$\mathrm{P}\big(\mu - \tfrac{1}{2}\sigma \le X \le \mu + \sigma\big) = \mathrm{P}(X \le \mu + \sigma) - \mathrm{P}\big(X < \mu - \tfrac{1}{2}\sigma\big) = a - (1-b) = a + b - 1.$$

조건부확률(conditional probability)의 경우, 분자의 사건은 $\{\mu - \frac{1}{2}\sigma \le X \le \mu + \frac{1}{2}\sigma\}$이고 그 확률은 $b - (1-b) = 2b - 1$이며, 한편 $\mathrm{P}(X \ge \mu - \frac{1}{2}\sigma) = 1 - (1-b) = b$이다. 따라서

$$\mathrm{P}\big(X \le \mu + \tfrac{1}{2}\sigma \mid X \ge \mu - \tfrac{1}{2}\sigma\big) = \frac{2b-1}{b}.$$

소문항 (ii)(a)

우유 팩이 탈지유(skimmed milk)인 사건을 S, 전지유(full-fat milk)인 사건을 F라 하면 $\mathrm{P}(S) = 0.6$이고 $\mathrm{P}(F) = 0.4$이다. 부피(ml)를 V라 하자.

> **생각해 보기** 각 우유 종류에 대하여 500 ml와 505 ml는 σ 단위로 어디에 놓이는가?

탈지유($\mu = 500, \sigma = 10$)에서 500은 μ이고 505는 $\mu + \frac{1}{2}\sigma$이다. 전지유($\mu = 495, \sigma = 10$)에서 500은 $\mu + \frac{1}{2}\sigma$이고 505는 $\mu + \sigma$이다. 띠에서 읽어 내면

$$\mathrm{P}(V < 505 \mid S) = b, \qquad \mathrm{P}(500 < V < 505 \mid S) = b - \tfrac{1}{2},$$
$$\mathrm{P}(V < 505 \mid F) = a, \qquad \mathrm{P}(500 < V < 505 \mid F) = a - b.$$

전확률 법칙(law of total probability)에 의하여

$$\mathrm{P}(V < 505) = 0.6b + 0.4a,$$
$$\mathrm{P}(500 < V < 505) = 0.6\big(b - \tfrac{1}{2}\big) + 0.4(a-b) = 0.4a + 0.2b - 0.3.$$

사건 "500 ml보다 많고 505 ml보다 적다"가 바로 우리가 원하는 교집합이므로

$$\mathrm{P}(V > 500 \mid V < 505) = \frac{0.4a + 0.2b - 0.3}{0.4a + 0.6b} = \frac{4a + 2b - 3}{4a + 6b}.$$

소문항 (ii)(b)

무작위로 고른 우유 팩의 개수가 매우 많으면, 관측된 비율은 해당 확률의 추정값이 된다. 첫 번째 정보는 $\mathrm{P}(V \le 505) = 0.7$이라는 것이고, 이는 소문항 (a)의 계산에 의하여

$$0.6b + 0.4a = 0.7, \quad \text{즉} \quad 4a + 6b = 7. \tag{1}$$

? 생각해 보기 각 우유 종류에 대하여 $\mathrm{P}(V \geq 495)$는 얼마인가?

탈지유에서 495는 $\mu - \frac{1}{2}\sigma$이므로 $\mathrm{P}(V \geq 495 \mid S) = b$이다(이는 소문항 (i)의 분모였다). 전지유에서 495는 μ이므로 대칭성에 의하여 $\mathrm{P}(V \geq 495 \mid F) = \frac{1}{2}$이다. 베이즈 정리(Bayes' theorem)에 의하여

$$\mathrm{P}(F \mid V \geq 495) = \frac{0.4 \times \frac{1}{2}}{0.6b + 0.4 \times \frac{1}{2}} = \frac{0.2}{0.6b + 0.2}.$$

이것을 $\frac{1}{3}$과 같게 놓으면 $0.6b + 0.2 = 0.6$이므로 $b = \frac{2}{3}$이다. 그러면 (1)에서 $4a = 7 - 4 = 3$이므로

$$a \approx \tfrac{3}{4}, \qquad b \approx \tfrac{2}{3}.$$

관측된 비율은 참된 확률의 근삿값일 뿐이므로, 이 값들은 추정값에 불과하다.

검산

- 추정값이 타당하다. $\mu < \mu + \frac{1}{2}\sigma < \mu + \sigma$이므로 반드시 $\frac{1}{2} < b < a < 1$이어야 하는데, 실제로 그렇다.
- 다시 대입하면 $0.6 \cdot \frac{2}{3} + 0.4 \cdot \frac{3}{4} = 0.4 + 0.3 = 0.7$이고, $0.2/(0.4 + 0.2) = \frac{1}{3}$이다.
- 정규분포(Normal distribution)에서는 $a \approx 0.841$, $b \approx 0.691$이고, 이 값들로 소문항 (i)을 계산하면 0.533과 0.553을 얻는다. (정규분포가 아닌) 라플라스 분포(Laplace distribution)를 10^5개 이상의 표본으로 시뮬레이션한 결과는 소문항 (i)과 (ii)(a)의 모든 공식과 일치했다. 예를 들어 (ii)(a)에서는 공식의 값 0.2515에 대하여 0.2518을 얻었다.

흔한 실수

- $1 - b$ 대신 $\mathrm{P}(X \leq \mu - \frac{1}{2}\sigma) = b$라고 쓰는 것. 대칭성이 뒤집는 것은 누적확률이 아니라 꼬리이다.
- 두 우유 종류의 평균(mean)이 서로 다르다는 것을 잊는 것. 그래서 같은 부피(예를 들어 500 ml)라도 각 종류마다 평균으로부터 떨어진 표준편차(standard deviation)의 개수가 다르다.
- 두 종류 모두에 $\mathrm{P}(V \geq 495) = b$를 쓰는 것. 전지유에서 495 ml는 평균이다.

STEP 너머

이 문제의 요점은 분포를 전혀 모르더라도 대칭성과 전확률 법칙만으로 꽤 멀리 갈 수 있다는 것이다. 여기서 두 모집단을 섞은 것은 간단한 **혼합 모형**(*mixture model*)이며, 소문항 (ii)(b)는 데이터로부터 그 모수(parameter)를 추정하는 일의 초보적인 형태이다. 이와 관련된 분포에 의존하지 않는 결과로 체비쇼프 부등식(Chebyshev's inequality)이 있는데, 이는 분산(variance)만을 사용하여 $\mathrm{P}(|X-\mu| \geq k\sigma)$의 상계를 준다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문), 또는 Grimmett과 Welsh의 *Probability: An Introduction*(확률: 입문)을 참고하라.

Beyond STEP 풀이: https://beyondstep.keithmcnulty.org/ko/. 풀이 글은 CC0 1.0이다. 문제 자체의 저작권은 © Cambridge University Press & Assessment에 있으며 여기에 싣지 않는다. 영어 원문을 기계 번역하였으며 아직 검토되지 않았다.